BIOMETRICS 57, 240-244

Minimising model fitting objectives that contain spurious local minima by bootstrap restarting

S. N. Wood
The Mathematical Institute, University of St Andrews
North Haugh, St Andrews KY16 9SS, U.K.
email: snw@st-and.ac.uk

SUMMARY. Objective functions that arise when fitting non-linear models often contain local minima that are of little significance except for their propensity to trap minimisation algorithms. The standard methods for attempting to deal with this problem treat the objective function as fixed and employ stochastic minimisation approaches, in the hope of randomly jumping out of local minima. This note suggests a simple trick for performing such minimisations which can be employed in conjunction with most conventional non-stochastic fitting methods. The trick is to stochastically perturb the objective function, by bootstrapping the data to be fit. Each bootstrap objective shares the large scale structure of the original objective, but has different small scale structure. Minimisations of bootstrap objective functions are alternated with minimisations of the original objective function starting from the parameter values with which minimisation of the previous bootstrap objective terminated. An example is presented, fitting a non-linear population dynamic model to population dynamic data and including a comparison of the suggested method with simulated annealing. Convergence diagnostics are discussed.

KEY WORDS: Non-linear model fitting; Simulated annealing; Stochastic optimization; Global optimization; Population dynamic model fitting; Ecological model.

1. Introduction

Most scientific process models are non-linear, their structure and complexity being determined by the mechanisms which they describe, rather than considerations of their convenience for statistical work. Attempts to use such models for statistical inference are often hampered by the difficulty of fitting them. Dynamic models are particularly taxing. Apparently simple dynamic equations often show a bewildering range of complex behaviours as parameters are altered (e.g. May, 1976; May and Oster, 1976; Murray, 1989; Gurney and Nisbet, 1998). This behaviour can translate into fitting objectives that contain much fine structure, including local minima, which are of a scale below that of the sampling induced uncertainty in the objective. Despite this fact that many features of difficult fitting objectives are nuisance features, the standard minimisation methods for such cases treat the objective function as fixed, while using a variety of stochastic strategies to avoid or escape from local minima. The most common approaches are either to use non-stochastic optimization methods started from a large number of randomly chosen initial parameters, to use simulated annealing, or to employ ad hoc problem specific tactics.

The difficulties with the random starts approach are two- fold. One must define a sensible region of parameter space from which to start the process of fitting, but often the sub- region that corresponds to remotely useful model fits is relatively small, so that most of the starting values are effectively wasted. The second problem is less obvious. Consider an objective made up of a large sloping depression densely pitted with shallow localised minima. Randomly chosen parameters will almost certainly be a considerable distance and some way uphill from the overall minimum. Hence, the overall minimum can only be attained by accidentally avoiding all the local minima between it and the

starting parameters. In other words, the random starting points are an attempt to locate, by chance, a point in parameter space that allows a clear run to the function minimum without interruption by local minima. This approach is rarely likely to be efficient, and becomes ever less so with increasing dimensionality.

Simulated annealing is a preferable approach to random starts (Brooks and Morgan, 1995, 1994). The basic concept is as follows: (i) a trial parameter vector is updated according to some rule; (ii) if the new parameters are an improvement over the old ones then they are always accepted, but if worse they are accepted with a probability that depends on how much worse, and how far minimisation has progressed; (iii) uphill steps of a given size are made progressively less probable as the minimisation proceeds. The scope for variations on the basic concept is obviously quite wide. For each problem, the user has to choose the region of parameter space to search, the way in which step up probabilities should be reduced, and the method to be used to generate trial steps.

The difficulty with all methods that treat the objective function as fixed, while escaping local minima by random jumps, is the difficulty of choosing the direction and size of the jumps. The information available at a local minimum is always useless in this respect: gradient and curvature within the immediate vicinity of a local minimum provide no guide as to the overall shape of the objective, or the direction of the global minimum.

2. Bootstrap Restart Optimization

Consider a fitting objective $f(\mathbf{p}, \mathbf{y})$ which is a function of data $\mathbf{y} = (y_1, y_2 \ldots y_n)'$ and parameters $\mathbf{p} = (p_1, p_2 \ldots, p_m)'$. Suppose also that you have a method capable of finding a local minimum of f with respect to $\mathbf{p}$, given starting parameters and a data vector. The bootstrap restarting approach is very simple:

1. Given a starting vector $\mathbf{p}_0$, find parameters which are at a minimum of $f(\mathbf{p}, \mathbf{y})$: $\hat{\mathbf{p}}_0$.
2. Repeat steps 3-5 for $i = 1, \ldots, k$.
3. Create a non- parametric or parametric bootstrap resample $\mathbf{y}_i^*$. From starting parameters $\hat{\mathbf{p}}_{i-1}$ find parameters which are at a minimum of $f(\mathbf{p}, \mathbf{y}_i^*)$: $\mathbf{p}_i^*$.
4. From starting parameters $\mathbf{p}_i^*$, find parameters that are at a minimum of $f(\mathbf{p}, \mathbf{y})$: $\mathbf{p}_i$.
5. If $f(\mathbf{p}_i, \mathbf{y}) \leq f(\hat{\mathbf{p}}_{i-1}, \mathbf{y})$ set $\hat{\mathbf{p}}_i = \mathbf{p}_i$ otherwise set $\hat{\mathbf{p}}_i = \hat{\mathbf{p}}_{i-1}$

$\hat{\mathbf{p}}_k$ contains the best fit parameters after k iterations. The idea is that although $f(\mathbf{p}, \mathbf{y}^*)$ will usually preserve the large scale features of $f(\mathbf{p}, \mathbf{y})$, small scale detail capable of trapping minimization methods will differ. Hence the method provides a way of escaping statistically spurious local minima in a way that automatically takes account of the large scale structure of the objective. The approach has the advantages of extreme simplicity and the fact that it can be used with most existing minimisation methods. Methods based on the local shape of the objective (e.g. gradient descent, Newton type methods, and even the polytope algorithm: see e.g Gill, Murray and Wright, 1981) will make good progress when outside local minima. When such a method gets stuck, it will be freed at the next bootstrap replicate not sharing the trapping minimum.

Two types of bootstrapping are suggested above in step 3. In this context "non-parametric" bootstrapping is sampling with replacement from the data to be fitted, so that any covariates are resampled with the datum that they relate to (sometimes known as "case resampling"): each resample would normally be of the same size as the original data set, although in some circumstances it may be appropriate to take smaller resamples in order to increase the perturbation of the objective function. When this approach is impractical (for example in some time-series problems) "parametric" bootstrapping can be used. Here, this means taking the fitted values predicted by the best fit model so far, and perturbing these with pseudo-random deviates generated from the assumed sampling distribution of the data (again using the best fit parameters so far, and estimating any remaining scale parameters by moment estimators). The parametric approach is more problematic than the non-parametric method. In particular it requires a reasonable initial model fit in order to be able to generate plausible bootstrap resamples. Of course, some circumstances might require more elaborate bootstrapping schemes than the two suggested here: bootstrapping methodology is covered extensively in Efron and Tibshirani (1993) and Davison and Hinkley (1997) and an alternative bootstrapping method for dealing with multiple local minima is suggested in Tibshirani and Knight (1999).

3. Convergence

Convergence of stochastic optimization methods is usually difficult to diagnose with certainty, and ascertaining whether a non-linear optimization method has located a global optimum is usually impossible. Hence it will not usually be possible to guarantee that the bootstrap restarting procedure has located a global optimum. Instead there are a number of informal diagnostics that can be examined:

1. Model fitting can be repeated from substantially different initial parameter values $\mathbf{p}_0$, to check that the same best fit parameters $\hat{\mathbf{p}}_k$ are identified from each. If not then k should be increased.

2. $f(\hat{\mathbf{p}}_i, \mathbf{y})$ can be plotted against i. If this plot shows no sign of levelling off, then it is clear that a global optimum has not yet been reached.

3. Let $\mathbf{p}_i$ be the j^{th} distinct local minima discovered. A plot of j against i should level off once the global minimum has been located, as the bootstrap restarting procedure ceases to find new local minima, and simply re-visits old ones.

Clearly none of these diagnostics is conclusive, and some combination of them should be used. Figure 3 provides two convergence diagnostic plots for the example presented in the next section. Note also that the decision about whether or not a local minimum is distinct from all previous local minima is not always easy: there can be difficulties if parameters are nearly co-linear so that the objective is locally nearly flat in some direction(s), or if relatively sloppy convergence criteria are being used for the optimization method. The latter problem arises when computational speed is important - considering only the optimization problem, there is little point in expending effort on precisely identifying the exact location of local minima which will only be abandoned subsequently. This suggests using fairly relaxed convergence criteria at any one optimization step, but this in turn causes problems in using diagnostics that rely on being able to identify when a minimum is distinct from other minima.

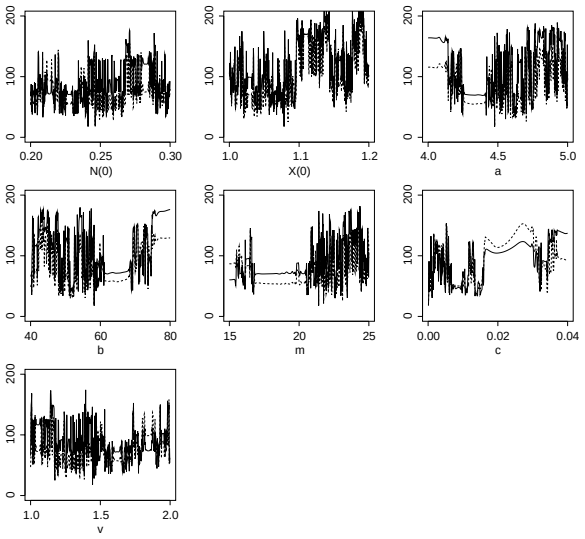

Figure 1: The solid lines show cross sections through the seven dimensional fitting objective $f(\mathbf{p}, \mathbf{y})$ for the model of section 4. The dashed lines are equivalent cross sections for an objective $f(\mathbf{p}, \mathbf{y}^*)$ based on a non-parametric bootstrap resample of the data. The transects were obtained by fixing all but one parameter at the minimum of $f(\mathbf{p}, \mathbf{y})$ and varying the remaining parameter.

4. Example: forest insect population dynamics

In this section I present a particularly taxing model fitting problem which arose from work aimed at understanding the mechanisms that cause forest insect cycles. The example provides an opportunity to compare the bootstrap restarting method with simulated annealing as described by Brooks and Morgan (1995).

The Pine Looper Moth (*Bupalus piniaria*) has caterpillars which feed on the foliage of Scots Pine (*Pinus sylvestris*) and can cause severe damage when sufficiently abundant. After an outbreak at Cannock Forest (in the North of England), so severe that a substantial area of forest had to be clear cut, the U.K. forestry commission set up an annual monitoring program for these insects (see Barbour, 1981, 1988, 1990; Broekhuizen, 1991). At Tentsmuir forest in Fife, Scotland the resulting data show pronounced regular cycles: it is not understood why, although there are a number of competing mechanistic explanations (see, for example, Broekhuizen, 1991, 1994). A host- parasitoid interaction (see e.g. Hassell, 1978), a consequence of inherited maternal quality (Ginzburg and Taneyhill, 1994), feedbacks between insect abundance and

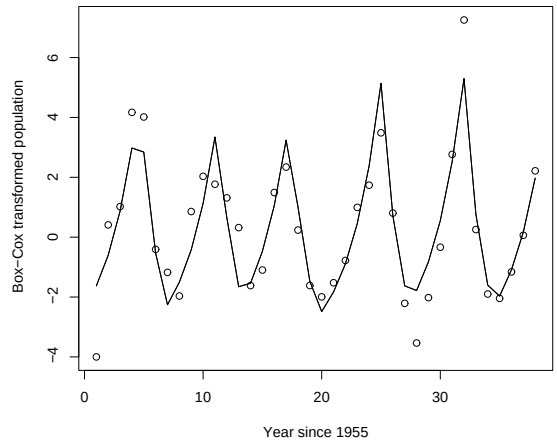

Figure 2: The best fit model trajectory found by fitting with bootstrap restarts (joined by solid line) and data (symbols) on the Box-Cox transformed scale (parameter 0.25).

plant secondary compounds, or simple consumer-resource interactions are some of the hypothesised causes of the cycle. Models implementing each of these mechanisms proved difficult to fit by conventional optimization methods alone, but bootstrap restarting proved effective in all cases. The most difficult model to fit was a simple description of the interaction between the caterpillars and their food supply, which produces a fitting objective function that looks like a suitable candidate for simulated annealing (see figure 1). I will concentrate on this model, and use it to compare bootstrap restarting and simulated annealing.

Let X_t be the quantity (suitably scaled) of available food (pine needles) at beginning of year t. N_t is the population of insects at the time of census in year t. The model is based on the assumption that the most recent three years growth of needles contribute to the food supply, each year class of needles being reduced by the insects that consume it. It is further assumed that on average the trees produce the same number of new needles each year. Insect survival and reproduction is assumed to be directly dependent on food supply, so that the net annual reproductive rate of the insects is a saturating function of food abundance. The model used is as follows:

$$N_{t+1} = N_t \frac{aX_t^m}{b + X_t^m} \quad (1)$$

$$X_{t+1} = 1 + P(N_t) + P(N_t)P(N_{t-1}) \quad (2)$$

$$\text{where } P(N) = c + (1-c)e^{-N^2/v} \quad (3)$$

$P(N)$ is the proportion of needles surviving from one year to the next given a population N of insects. The parameters a , b, m, c and v must all be estimated by fitting, along with starting conditions $X(0)$ and $N(0)$. The available data are estimates of N_t (for full details see Barbour, 1981, 1988, 1990; Broekhuizen 1991), and the equation for X_t has been re-parameterized to remove a scale parameter that can not be estimated without direct observations of food abundance. Simple models of this sort can show remarkably subtle behaviour. For example, if the model cycle period (defined, for example, as the mean time between peak abundances) is non-integer then the time between visits to exactly the same population value can become very long (or infinite): such model trajectories can display a good deal of small scale variability from cycle to cycle leading to shallow and localised minima in a fitting objective function. Furthermore, the cycle period does not necessarily change smoothly with parameters. There can be sharp changes between discrete frequencies, tending to produce fitting objectives with many local minima each shallower than the sampling uncertainty associated with the objective function.

For illustrative purposes, I will consider only the simple fitting objective:

$$f(\mathbf{p}, \mathbf{y}) = \frac{1}{2}\sum_{i=1}^{n}(\tilde{y}_i - \tilde{m}_i(\mathbf{p}))^2$$

where $\mathbf{p}$ is the vector of model parameters, y_i is the i^{th} observation of insect density, and I have written m_i for the model prediction of that density. $\tilde{x}_i$ denotes the Box-Cox transformation of x_i (i.e. $\tilde{x}_i = (x_i^\lambda - 1)/\lambda$. λ was set to 0.25, a value chosen by searching for the best normal scores residual plot). Starting values for the minimisation were obtained by the expedient of augmenting the objective with a least squares term $c\sum_{i=1}^{10}(ACF(\tilde{\mathbf{m}})_i - ACF(\tilde{\mathbf{y}})_i)^2$, where $ACF(\mathbf{x})_i$ is the i^{th} term of the auto-correlation function of series $\mathbf{x}$, and c is a weight chosen to give adequate weight to the ACF part of the objective. This extra least squares term was dropped once rough starting values were found (i.e. c was set to zero).

Figure 1 displays cross-sections through $f(\mathbf{p}, \mathbf{y})$ with equivalent cross-sections through a single $f(\mathbf{p}, \mathbf{y}^*)$ superimposed. The plots show transects through the function minimum obtained by varying single parameters. Notice that although the original and bootstrapped objectives have similar shape, they differ in detail. Note also that the objective looks a little worse than it is - minima in a

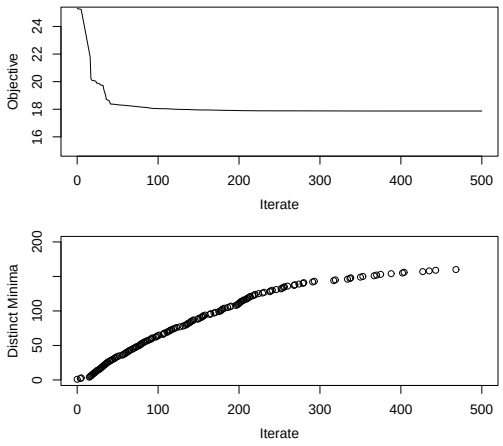

Figure 3: Convergence diagnostic plots for the bootstrap restarting model fit reported in section 4. The upper plot shows $f(\hat{\mathbf{p}}_i, \mathbf{y})$ against i. The lower plot shows the number of distinct local minima discovered against iterate (symbols are plotted at each iterate at which a new local minimum was discovered). At convergence both plots should level off.

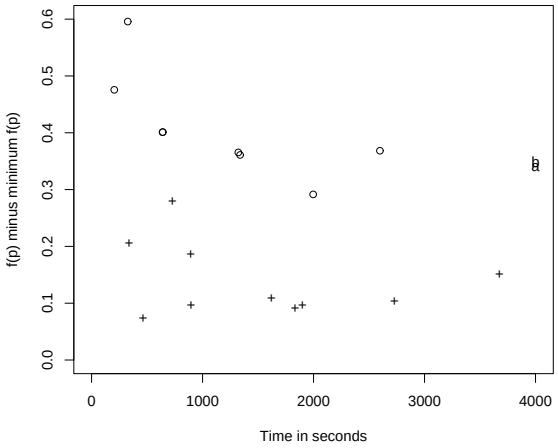

Figure 4: Plots of the value of the 'best fit' objective function less the overall best fit found by the bootstrap restart method, against the average time taken for this minimisation. Times and objective values are averages over 10 replicates of each fitting method with each set of method controlling parameters. (+) denotes bootstrap restarting method, while (o) denotes simulated annealing. (a) and (b) are simulated annealing points for which time should be 9297 and 6173, respectively.

one dimensional transect are not necessarily minima in the full 7 dimensional objective. Figure 2 shows the best fit obtained when minimising the fitting objective by Quasi-Newton (see Gill et al., 1981 and Gill et al., 1974 for a stable implementation or Press et al., 1992 for a less stable but simpler one) with 500 bootstrap restarts. Figure 3 provides two convergence diagnostic plots for this model fit.

I compared the performance of bootstrap restarting with simulated annealing as described in Brooks and Morgan (1995). The method requires that the user supply a bounding region within which the parameters lie, as well as a coefficient controlling the annealing schedule and a coefficient controlling how many parameter vectors are tried at each annealing temperature. It is also necessary to supply a starting temperature. I added a final Quasi-Newton step at the end of the simulated annealing to ensure that the method ended up in a minimum of some sort. I chose the bounding region of parameter space to be a box centred on the best fit parameters found by bootstrap restarting, with the distance to the box faces defined by the differences between the starting parameters for the bootstrap restarting and the best fit parameters. Both methods performed reasonably well, but despite considerable experimentation and some very long simulated annealing runs, the simulated annealing minimum was consistently higher than the minimum achieved by bootstrap restarting.

The currency that is probably of most interest for comparing methods, combines the quality of the achieved minimum with the amount of effort taken to reach it. I measured quality of each estimated minimum by the difference between the minimum and the lowest minimum achieved overall (17.68), and measured effort in seconds of computer time taken (using a Pentium II 400Mhz with Windows NT and 32-bit code). Figure 4 plots these two measures against each other for Bootstrap restarting (+) and Simulated Annealing (o). Each point is the mean of 10 replicates employing the same controlling parameters. The better the method the closer its points should be to zero on the vertical axis, and the more rapidly zero should be approached with increasing effort (time). Clearly, while the difference in performance is not huge, the bootstrap restarting method is performing consistently better than simulated annealing, in this case.

5. Discussion

The bootstrap restarting method is designed to deal with model fitting problems that are made difficult by the presence of small scale 'nuisance' features in the fitting objective. Within this context, there are three features of the method that suggest that it should be useful. Firstly it is very simple, and is therefore easy to implement given any conventional optimization method. Secondly, it provides

a complementary approach to traditional methods for difficult objective functions, which treat the objective function itself as fixed. Thirdly, the example suggests that bootstrap restarting may be quite effective even for fitting problems for which a simulated annealing method at first sight appears to provide the only hope.

Relative to an optimization method used without bootstrap restarting, the bootstrap restarting method will never produce a worse fit, and will often improve fit. However, this is very far from guaranteeing that a globally optimum fit will be achieved, even as the number of iterates tends to infinity, and, as is usual with stochastic optimization schemes, there is currently no alternative to the use of informal diagnostics for judging convergence.

Acknowledgements

Thanks to Bruce Kendall for providing the original problem and discussion of the biological example presented here and to the referees for helpful comments and suggestions on an earlier draft of the paper.

References

Broekhuizen, N. (1991). The Population Dynamics of the Pine Looper Moth, *Bupalus piniaria* L. (Ledpidoptera: Geometridae). Unpublished Phd thesis, Imperial College, London.

Broekhuizen, N., Hassell, M.P. and Evans, H.F. (1994). Common mechanisms underlying contrasting dynamics in 2 populations of the Pine Looper Moth. *Journal of Animal Ecology* **63**, 245-255.

Barbour, D.A. (1981). Population Dynamics of the Pine Looper Moth, *B. piniaria* (L.) (Lepidoptera, Geometridae) in British Pine Forests. Unpublished PhD thesis, University of Edinburgh.

Barbour, D.A. (1988). The Pine Looper in Britain and Europe. In *Dynamics of Forest Insect Populations: Patterns, Causes, Implications*, A.A. Berryman (ed), 291-308. New York: Plenum Press.

Barbour, D.A. (1990). Synchronous fluctuations in spatially separated populations of cyclic forest insects. In *Population Dynamics of Forest Insects*, A.D. Watt, S.R. Leather, M.D. Hunter and N.A.C. Kidd (eds), 339-346. Andover, Hants: Intercept.

Brooks, S.P. and Morgan, B.J.T. (1994). Automatic starting point selection for function optimization. *Statistical Computation* **4**, 173-177.

Brooks, S.P. and Morgan, B.J.T. (1995). Optimization using simulated annealing. *The Statistician* **44**, 241-257.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Gill, P.E., Golub, G.H., Murray, W. and Saunders, M.A. (1974). Methods for modifying Matrix Factorizations. *Mathematics of Computation* **28**, 505-535.

Gill, P.E., Murray, W. and Wright, M.H. (1981). *Practical Optimization*. London: Academic Press.

Ginzburg, L.R. and Taneyhill, D.E. (1994). Population cycles of forest lepidoptera - A maternal effect hypothesis. *Journal of Animal Ecology* **63**, 79-92.

Gurney, W.S.C. and Nisbet, R.M. (1998). *Ecological Dynamics*. New York: Oxford University Press.

Hassell, M.P. (1978). *The dynamics of Arthropod Predator-Prey Systems*. Princeton: Princeton University Press.

May, R.M. (1976). Simple mathematical models with very complicated dynamics. *Nature* **261**, 459-467.

May, R.M. and Oster, G.F. (1976). Bifurcations and dynamic complexity in simple ecological models. *American Naturalist* **110**, 573-599.

Murray, J.D. (1989). *Mathematical Biology*. Heidelberg: Springer-Verlag.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992). *Numerical Recipes in C*. Cambridge: Cambridge University Press.

Tibshirani, R. and Knight, K. (1999). Model search by bootstrap bumping. *Journal of Computational and Graphical Statistics* **8**, 671-686.